

# Task Decoupled Framework for Reference-based Super-Resolution - Supplementary Material

Yixuan Huang<sup>1\*</sup>, Xiaoyun Zhang<sup>1\*†</sup>, Yu Fu<sup>1</sup>, Siheng Chen<sup>1,2</sup>, Ya Zhang<sup>1,2</sup>, Yanfeng Wang<sup>1,2</sup><sup>†</sup>, Dazhi He<sup>1</sup>

<sup>1</sup>Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, <sup>2</sup>Shanghai AI Laboratory

{huangyixuan, xiaoyun.zhang, fyu11, sihengc, ya\_zhang, wangyanfeng, hedazhi}@sjtu.edu.cn

In the supplementary file, we will explain some details of the network structures in Section.1. In Section.2, we introduce the training loss of the texture transfer model. And the robustness to irrelevant references will be further analyzed in Section.3. The details of confidence map  $C$  and the residual reconstruction loss are discussed in Section.4. We also analyze the computation cost and limitations in Section.5 and Section.6. Finally, we provide more visual results of the reference-underuse issue and the reference-misuse issue in Section.7 and more visual comparisons with state-of-the-art methods in Section.8.

## 1. Network Structures

We use 23 basic blocks of RRDB [5] in the Single Image Super-Resolution(SISR) network. As for the texture extraction module, We adopt the pre-trained contrastive correspondence network of  $C^2$ -Matching [2] as our feature extractor, because of its explicit robust matching. And we use eight residual blocks in the texture transfer module.

## 2. Loss Functions

**Reconstruction loss.** For the final output image  $I_{RefSR}$  We adopt  $L_1$  loss as the reconstruction loss as

$$\mathcal{L}_{rec} = \|\mathbf{I}_{HR} - \mathbf{I}_{RefSR}\|_1, \quad (1)$$

where  $\mathbf{I}_{HR}$  denotes the ground truth image. For the texture output  $I_{Tex}$ , we adopt the residual of  $\mathbf{I}_{HR}$  and the SISR output image  $\mathbf{I}_{SISR}$  as supervision:

$$\mathcal{L}_{rec}^{Tex} = \|\mathbf{I}_{HR} - (\mathbf{I}_{SISR} + \mathbf{I}_{Tex})\|_1, \quad (2)$$

**Perceptual loss.** The perceptual loss [3] is expressed as

$$\mathcal{L}_{per} = \|\phi_i(\mathbf{I}_{HR}) - \phi_i(\mathbf{I}_{RefSR})\|_2, \quad (3)$$

where  $\phi_i$  denotes the  $i$ -th layer features of VGG19 model. Here we use the relu5.1 features.

\*Equal contribution(co-first authors).

†Corresponding author.

Table 1. **Quantitative comparison on robustness to irrelevant reference.** Our method achieves the best performance when irrelevant references are given.

Method	CUFED5*	Urban100*
	PSNR/SSIM	PSNR/SSIM
TTSR [6]	26.40/0.782	25.85/0.785
MASA [4]	26.59/0.784	26.16/0.789
$C^2$ -Matching [2]	26.49/0.784	26.17/0.791
<b>Ours</b>	<b>26.83/0.793</b>	<b>26.91/0.809</b>

**Adversarial loss.** WGAN-GP [1] is employed to improve the visual quality. The adversarial loss  $L_{adv}$  can be interpreted as :

$$\mathcal{L}_D = D(\mathbf{I}_{RefSR}) - D(\mathbf{I}_{HR}) + \lambda(\nabla_{\hat{\mathbf{I}}} \|D(\hat{\mathbf{I}})\|_2)^2, \quad (4)$$

$$\mathcal{L}_G = -D(\mathbf{I}_{RefSR}), \quad (5)$$

**Full objective.** The full objective is defined as

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{rec}^{Tex} + \lambda_{per}\mathcal{L}_{per} + \lambda_{adv}\mathcal{L}_{adv}, \quad (6)$$

where the coefficients  $\lambda_{per}$  and  $\lambda_{adv}$  are  $10^{-2}$  and  $10^{-4}$ , respectively.

## 3. Robustness to Irrelevant Reference Images

To evaluate the robustness to irrelevant references, we build datasets CUFED5\* and Urban100\*. In CUFED5\*, We randomly select a reference image for each input image, ensuring the reference image is irrelevant to input image. So is Urban100\*. We compare our method with other three state-of-the-art RefSR methods, including TTSR [6], MASA [4] and  $C^2$ -Matching [2]. As shown in Table 1, our method achieves the best performance when incongruous references are given, indicating the superiority of our framework on robustness to irrelevant references.

Table 2. Ablation study of confidence map  $C$  and the residual reconstruction loss  $\mathcal{L}_{rec}^{Tex}$ .

Model	w/o $C$	w/o $\mathcal{L}_{rec}^{Tex}$	Full model
PSNR/SSIM	28.49/0.847	28.43/0.845	28.64/0.850



Figure 1. visualizations of confidence map and the output image of texture extraction module.

#### 4. Discussion of confidence map and the residual reconstruction loss

Our task decoupled framework can avoid the reference-misuse issue from the following two aspects. (1) First is the confidence map  $C$ . The confidence map has a low weight for the not-relevant region of the reference image, preventing the irrelevant texture deteriorating the output. (2) Second is the residual reconstruction loss  $\mathcal{L}_{rec}^{Tex}$  (Eq.11). Supervised by the ground-truth image, the texture extraction module is naturally capable of removing the irrelevant content. The ablation study in Table 3 show the effectiveness of the confidence map  $C$  and  $\mathcal{L}_{rec}^{Tex}$ . Fig.1 shows a higher degree of confidence(whiter) and more detailed texture in  $I_{Tex}$  in the case of related reference, compared with unrelated reference.

#### 5. Analysis of computation cost

To tackle two key issues pointed in the paper, we make a two-branch design to obtain a better performance, which decouples the super-resolution task and texture transfer task. As a result, compared to previous coupled RefSR methods, it is expected that our parameters and run-time increase, due to the extra SISR network. However, we can decrease the computation cost by using a lighter SISR model.

As shown in Table 3, we compare the parameters, runtime and performance with  $C^2$ -Matching. We train a light model of our framework (Ours-light). The data before '+' in the Parameters column means the parameters of the SISR network, while the data after '+' means the rest parameters. As we use a lighter SISR network, the computation cost decrease a lot with less performance reduction. It is a trade-off between computation cost and performance.

Table 3. Comparison of parameters and runtime between our method and  $C^2$ -Matching.

Model	Parameters	Runtime	PSNR/SSIM
$C^2$ -Matching	8.9M	72ms	28.24/0.841
Ours-light	272K+10.9M	94ms	28.46/0.846
Ours	15.9M+10.9M	123ms	28.64/0.850

#### 6. Limitations

Our method could potentially incur two limitations: (1) First is the increase of computation cost. Compared to previous coupled RefSR methods, our decoupled framework adds a SISR network branch, causing more computation cost. However, we can reduce the cost by using a lighter SISR network, and make a trade-off between computation cost and performance. (2) Another is a common issue for RefSR methods. That is, when the input has very low quality, it may deteriorate the alignment step in texture extraction module, resulting in an unsatisfactory final image.

#### 7. More Visual Results of the Reference-Underuse Issue and the Reference-Misuse Issue

We show more visual results of the reference-underuse issue and more comparisons of the Model-Both and the Model-Ref in Fig.2. Model-Ref transfers more detailed textures highly similar to GT. In Fig.3, we show more visual results of the reference-underuse issue and more comparisons of the Model-Both and the Model-LR. Model-LR eliminates the blur and artifacts of the Model-Both results.

#### 8. More Visual Comparisons with State-of-the-art Methods

We show more visual comparisons with other state-of-the-art methods, including RCAN [8], ESRGAN [5], RankSRGAN [7], SRNTT [9], TTSR [6], MASA [4] and  $C^2$ -Matching [2] in Fig.4 and Fig.5.  $C^2$ -Matching is the state-of-the-art RefSR method recently. Our results obtain more pleasing textures transferred from the reference image.

#### References

[1] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Adv. Neural Inform. Process. Syst.*, pages 5767–5777, 2017. 1

[2] Yuming Jiang, Kelvin C.K. Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via  $C^2$ -matching. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2103–2112, 2021. 1, 2

- [3] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Eur. Conf. Comput. Vis.*, pages 694–711, 2016. [1](#)
- [4] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: matching acceleration and spatial adaptation for reference-based image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6368–6377, 2021. [1](#), [2](#)
- [5] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Eur. Conf. Comput. Vis. Worksh.*, 2018. [1](#), [2](#)
- [6] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Bain-ing Guo. Learning texture transformer network for image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5791–5800, 2020. [1](#), [2](#)
- [7] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Int. Conf. Comput. Vis.*, pages 3096–3105, 2019. [2](#)
- [8] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Eur. Conf. Comput. Vis.*, pages 286–301, 2018. [2](#)
- [9] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7982–7991, 2019. [2](#)



Figure 2. **More visual results of the reference-underuse issue.** Model-Ref, which only use the aligned reference image to reconstruct the HR image, transfers more detailed textures from reference images when relevant high quality reference images are given. **(Zoom-in for best view)**

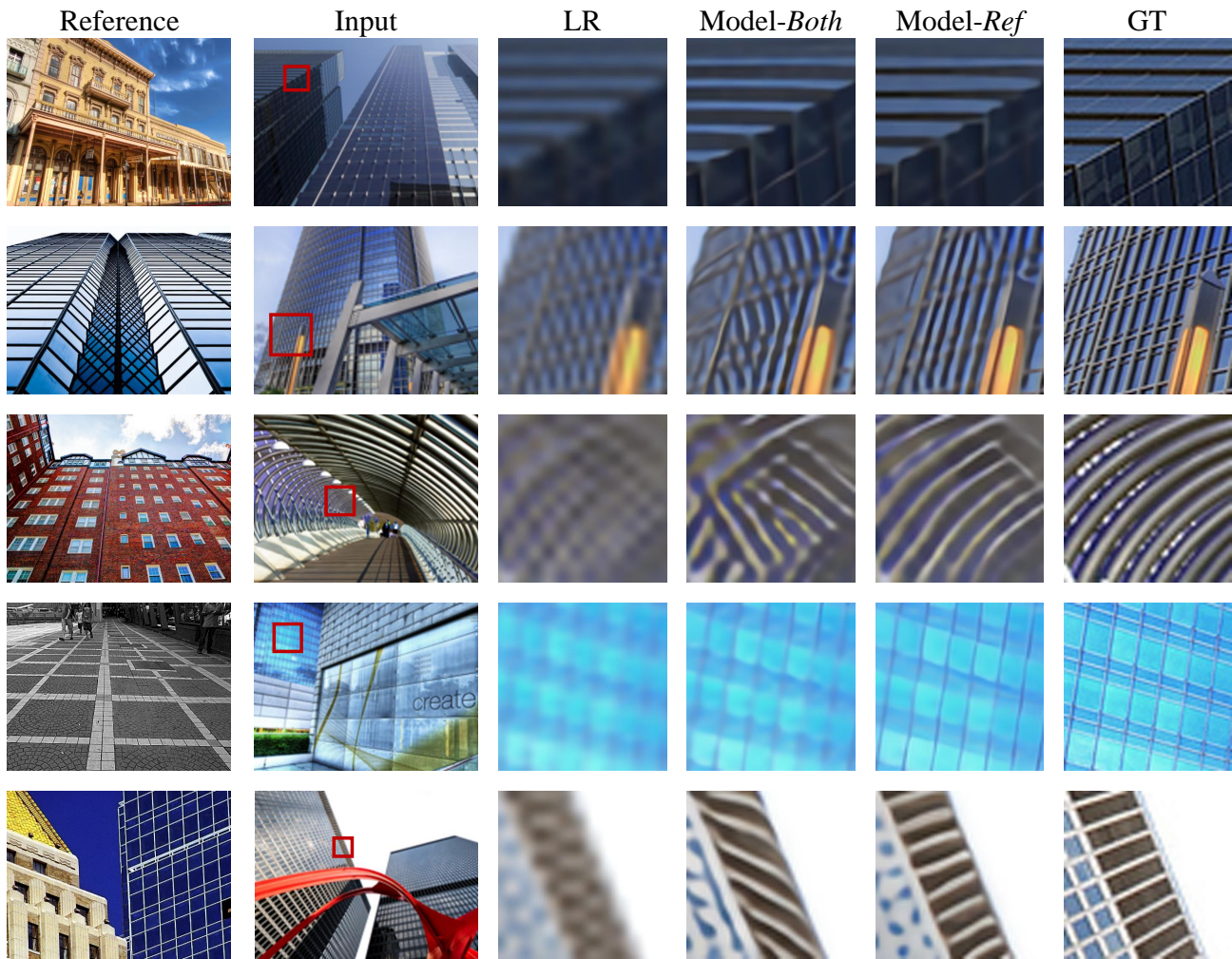


Figure 3. More visual results of the reference-misuse issue. Model-*LR*, which only use the input LR image to reconstruct the HR image, overcomes the blur and artifacts caused by reference interference when irrelevant reference images are given. (**Zoom-in for best view**)

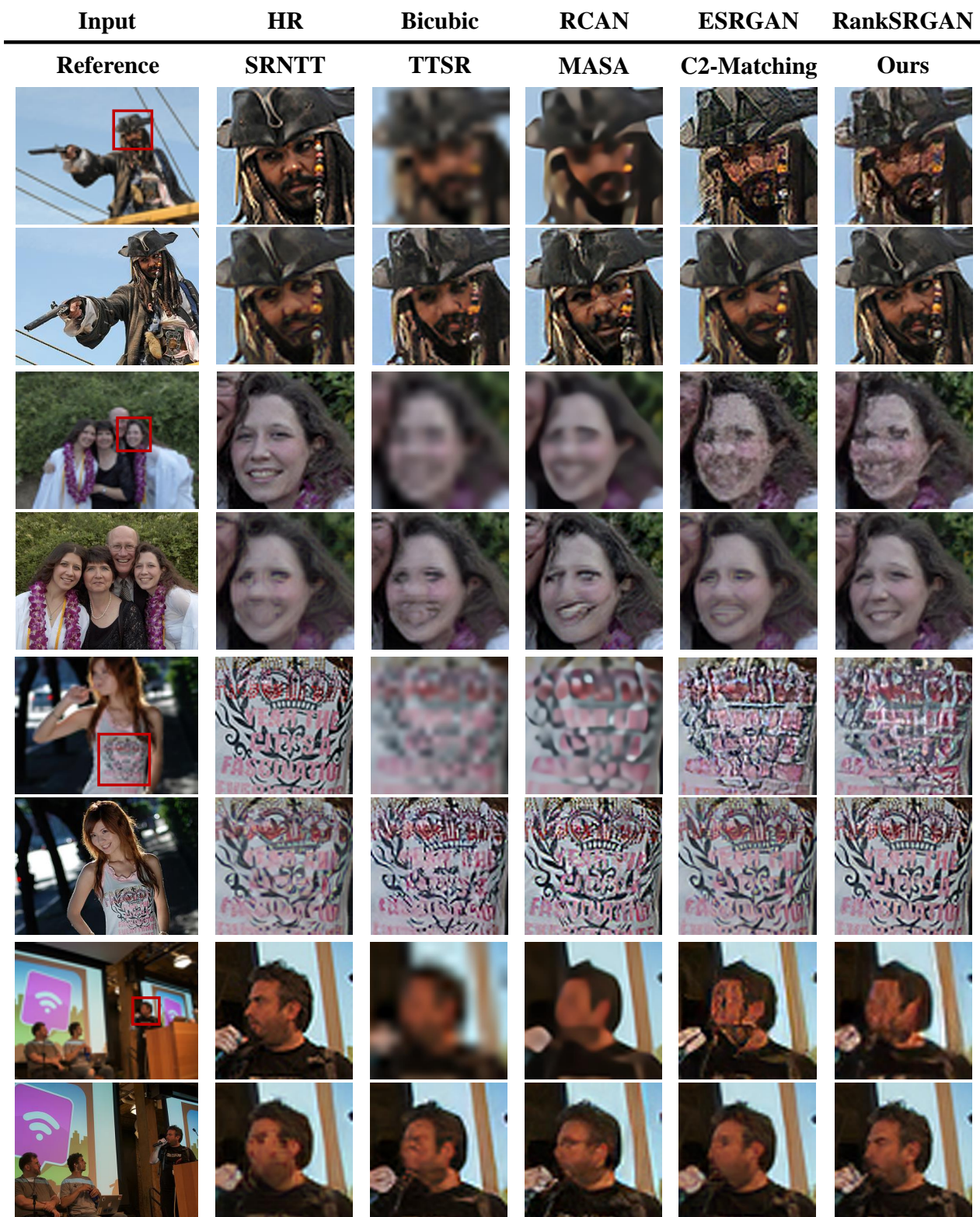


Figure 4. **More visual comparisons.** Only RCAN is trained with  $l_1$  loss, and others are all trained with GAN loss. Our method restores more realistic textures compared with other RefSR and SISR methods. (**Zoom-in for best view**)


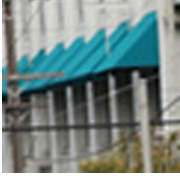
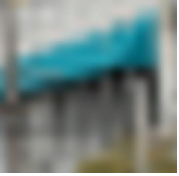

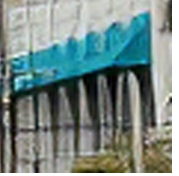







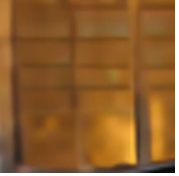





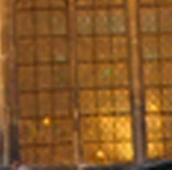

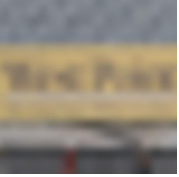
















Input	HR	Bicubic	RCAN	ESRGAN	RankSRGAN
Reference	SRNTT	TTSR	MASA	C2-Matching	Ours
					
					
					
					
					
					
					
					

Figure 5. **More visual comparisons.** Only RCAN is trained with  $l_1$  loss, and others are all trained with GAN loss. Our method restores more realistic textures compared with other RefSR and SISR methods. (**Zoom-in for best view**)